A Large-Scale Corpus for Conversation Disentanglement

Jonathan K. Kummerfeld[1*] **Sai R. Gouravajhala**[1] **Joseph J. Peper**[1]
Vignesh Athreya[1] **Chulaka Gunasekara**[2] **Jatin Ganhotra**[2]
Siva Sankalp Patel[2] **Lazaros Polymenakos**[2] **Walter S. Lasecki**[1]

Computer Science & Engineering[1]
University of Michigan

T.J. Watson Research Center[2]
IBM Research AI

Abstract

Disentangling conversations mixed together in a single stream of messages is a difficult task, made harder by the lack of large manually annotated datasets. We created a new dataset of 77,563 messages manually annotated with reply-structure graphs that both disentangle conversations and define internal conversation structure. Our dataset is 16 times larger than all previously released datasets combined, the first to include adjudication of annotation disagreements, and the first to include context. We use our data to re-examine prior work, in particular, finding that 80% of conversations in a widely used dialogue corpus are either missing messages or contain extra messages. Our manually-annotated data presents an opportunity to develop robust data-driven methods for conversation disentanglement, which will help advance dialogue research.

1 Introduction

When a group of people communicate in a common channel there are often multiple conversations occurring concurrently. Often there is no explicit structure identifying conversations or their structure, such as in Internet Relay Chat (IRC), Google Hangout, and comment sections on websites. Even when structure is provided it often has limited depth, such as threads in Slack, which provide one layer of branching. In all of these cases, conversations are *entangled*: all messages appear together, with no indication of separate conversations. Automatic disentanglement could be used to provide more interpretable results when searching over chat logs, and to help users understand what is happening when they join a channel. Over a decade of research has considered conversation disentanglement (Shen et al., 2006), but using datasets that are either small (2,500 messages, Elsner and Charniak, 2008) or not released (Adams and Martell, 2008).

We introduce a conversation disentanglement dataset of 77,563 messages of IRC manually annotated with reply-to relations between messages.[1] Our data is sampled from a technical support channel at 173 points in time between 2004 and 2018, providing a diverse set of speakers and topics, while remaining in a single domain. Our data is the first to include context, which differentiates messages that start a conversation from messages that are responding to an earlier point in time. We are also the first to adjudicate disagreements in disentanglement annotations, producing higher quality development and test sets. We also developed a simple model that is more effective than prior work, and showed that having diverse data makes it perform better and more consistently.

We also analyze prior disentanglement work. In particular, a recent approach from Lowe et al. (2015, 2017). By applying disentanglement to an enormous log of IRC messages, they developed a resource that has been widely used (over 315 citations), indicating the value of disentanglement in dialogue research. However, they lacked annotated data to evaluate the conversations produced by their method. We find that 20% of the conversations are completely right or a prefix of a true conversation; 58% are missing messages, 3% contain messages from other conversations, and 19% have both issues. As a result, systems trained on the data will not be learning from accurate human-human dialogues.

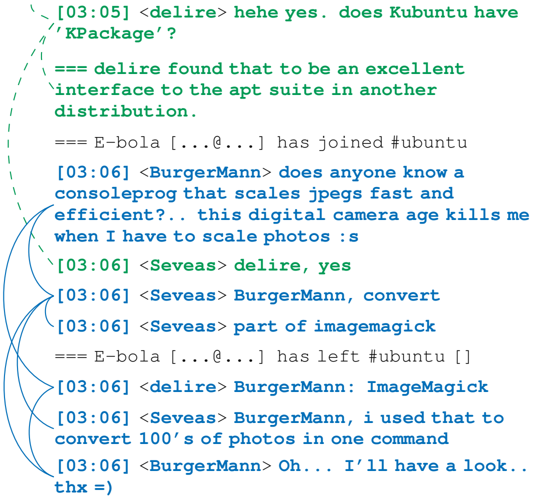
2 Task Definition

We consider a shared channel in which a group of people are communicating by sending messages that are visible to everyone. We label this data with a **graph** in which messages are nodes and edges indicate that one message is a response to another. Each connected component is a **conversation**.

[*] jkummerf@umich.edu

[1] https://jkk.name/irc-disentanglement

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3846–3856
Florence, Italy, July 28 - August 2, 2019.



```
[03:05] <delire> hehe yes. does Kubuntu have
'KPackage'?
=== delire found that to be an excellent
interface to the apt suite in another
distribution.
=== E-bola [...@...] has joined #ubuntu
[03:06] <BurgerMann> does anyone know a
consoleprog that scales jpegs fast and
efficient?.. this digital camera age kills me
when I have to scale photos :s
[03:06] <Seveas> delire, yes
[03:06] <Seveas> BurgerMann, convert
[03:06] <Seveas> part of imagemagick
=== E-bola [...@...] has left #ubuntu []
[03:06] <delire> BurgerMann: ImageMagick
[03:06] <Seveas> BurgerMann, i used that to
convert 100's of photos in one command
[03:06] <BurgerMann> Oh... I'll have a look..
thx =)
```

Figure 1: #Ubuntu IRC log sample, earliest message first. Curved lines are our *graph* annotations of reply structure, which define two *conversations* shown with blue solid edges and green dashed edges.

Figure 1 shows an example of two entangled conversations and their graph structure. It includes a message that receives multiple responses, when multiple people independently help BurgerMann, and the inverse, when the last message responds to multiple messages. We also see two of the users, delire and Seveas, simultaneously participating in two conversations. This multi-conversation participation is common.

The example also shows two aspects of IRC we will refer to later. **Directed** messages, an informal practice in which a participant is named in the message. These cues are useful for understanding the discussion, but only around 48% of messages have them. **System** messages, which indicate actions like users entering the channel. These all start with ===, but not all messages starting with === are system messages, as shown by the second message in Figure 1.

3 Related Work

IRC Disentanglement Data: The most significant work on conversation disentanglement is a line of papers developing data and models for the #Linux IRC channel (Elsner and Charniak, 2008; Elsner and Schudy, 2009; Elsner and Charniak, 2010, 2011). Until now, their dataset was the only publicly available set of messages with annotated conversations (partially re-annotated by Mehri and Carenini (2017) with reply-structure graphs), and has been used for training and evaluation in subsequent work (Wang and Oard, 2009; Mehri and Carenini, 2017; Jiang et al., 2018).

We are aware of three other IRC disentanglement datasets. First, Adams and Martell (2008) studied disentanglement and topic identification, but did not release their data. Second, Riou et al. (2015) annotated conversations and discourse relations in the #Ubuntu-fr channel (French Ubuntu support). Third, Lowe et al. (2015, 2017) heuristically extracted conversations from the #Ubuntu channel.[2] Their work opened up a new research opportunity by providing 930,000 disentangled conversations, and has already been the basis of many papers (315 citations), particularly on developing dialogue agents. This is far beyond the size of resources previously collected, even with crowdsourcing (Lasecki et al., 2013). Using our data we provide the first empirical evaluation of their method.

Other Disentanglement Data: IRC is not the only form of synchronous group conversation online. Other platforms with similar communication formats have been studied in settings such as classes (Wang et al., 2008; Dulceanu, 2016), support communities (Mayfield et al., 2012), and customer service (Du et al., 2017). Unfortunately, only one of these resources (Dulceanu, 2016) is available, possibly due to privacy concerns.

Another stream of research has used user-provided structure to get conversation labels (Shen et al., 2006; Domeniconi et al., 2016) and reply-to relations (Wang and Rosé, 2010; Wang et al., 2011a; Aumayr et al., 2011; Balali et al., 2013, 2014; Chen et al., 2017a). By removing these labels and mixing conversations they create a disentanglement problem. While convenient, this risks introducing a bias, as people write differently when explicit structure is defined, and only a few papers have released data (Abbott et al., 2016; Zhang et al., 2017; Louis and Cohen, 2015).

Models: Elsner and Charniak (2008) explored various message-pair feature sets and linear classifiers, combined with local and global inference methods. Their system is the only publicly released statistical model for disentanglement of chat conversation, but most of the other work cited above applied similar models. We evaluate their model on both our data and our re-annotated version of their data. Recent work has applied neural networks (Mehri and Carenini, 2017; Jiang et al.,

[2] This channel was first proposed as a useful data source by Uthus and Aha (2013a,b,c), who identified messages relevant to the Unity desktop environment, and whether questions can be answered by the channel bot alone.

Data Available?	Dataset		Messages	Parts	Part Length	Authors / part	Context	Anno. / msg
Yes	This work	Pilot	1,250	9	100–332 msg	19-48	0-100	1-5
		Train	47,500	95	500 msg	33-95	1000	1
			1,000	10	100 msg	20-43	1000	3+a
			18,963	48	1 hr	22-142	1000	1
		Dev	2,500	10	250 msg	76-167	1000	2+a
		Test	5,000	10	500 msg	79-221	1000	3+a
		Channel 2	2,600	1	5 hr	387	0	2+a
	Elsner and Charniak (2008)		2,500	1	5 hr	379	0	1-6
	Mehri and Carenini (2017)		530	1	1½ hr	54	0	3
Request	Riou et al. (2015)		1,429	2	12 / 60 hr	21/70	0	2/1
	Dulceanu (2016)		843	3	½–1½ hr	8-9	n/a	1
No	Shen et al. (2006)		1,645	16	35–381 msg	6-68	n/a	1
	Adams and Martell (2008)		19,925	38	67–831 msg	?	0	3
	Wang et al. (2008)		337	28	2–70 msg	?	n/a	1-2
	Mayfield et al. (2012)		?	45	1 hr	3-7	n/a	1
	Guo et al. (2017)		1,500	1	48 hr	5	n/a	2

Table 1: Annotated disentanglement dataset comparison. Our data is much larger than prior work, one of the only released sets, and the only one with context and adjudication. '+a' indicates there was an adjudication step to resolve disagreements. '?' indicates the value is not in the paper and the authors no longer have access to the data.

2018), with slight gains in performance.

Graph Structure: Within a conversation, we define a graph of reply-to relations. Almost all prior work with annotated graph structures has been for threaded web forums (Schuth et al., 2007; Kim et al., 2010; Wang et al., 2011b), which do not exhibit the disentanglement problem we explore. Studies that do consider graphs for disentanglement have used small datasets (Dulceanu, 2016; Mehri and Carenini, 2017) that are not always released (Wang et al., 2008; Guo et al., 2017).

4 Data

We introduce a manually annotated dataset of 77,563 messages: 74,963 from the #Ubuntu IRC channel,[3] and 2,600 messages from the #Linux IRC channel.[4] Annotating the #Linux data enables comparison with Elsner and Charniak (2008), while the #Ubuntu channel has over 34 million messages, making it an interesting large-scale resource for dialogue research. It also allows us to evaluate Lowe et al. (2015, 2017)'s widely used heuristically disentangled conversations.

When choosing samples we had to strike a balance between the number of samples and the size of each one. We sampled the training set in three ways: (1) 95 uniform length samples, (2) 10 smaller samples to check annotator agreement, and (3) 48 time spans of one hour that are diverse in terms of the number of messages, the number of participants, and what percentage of messages are directed. For additional details of the data selection process, see the supplementary material.

4.1 Dataset Comparison

Table 1 presents properties of our data and prior work on disentanglement in real-time chat.

Availability: Only one other dataset, annotated twice, has been publicly released, and two others were shared when we contacted the authors.

Scale: Our dataset is 31 times larger than almost any other dataset, the exception being one that was not released. As well as being larger, our data is also based on many different points in time. This is crucial because a single sample presents a biased view of the task. Having multiple samples also means our training and evaluation sets are from different points in time, preventing overfitting to specific users or topics of conversation.

Context: We are the first to consider the fact that IRC data is sampled from a continuous stream and the context prior to the sample is important. In prior work, a message with no antecedent could

[3] https://irclogs.ubuntu.com/

[4] From Elsner and Charniak (2008), including the 100 messages they did not annotate.

either be the start of a conversation or a response to a message that occurs prior to the sample.

Adjudication: Our labeling method is similar to prior work, but we are the first to perform adjudication of annotations. While some cases were ambiguous, often one option was clearly incorrect. By performing adjudication we can reduce these errors, creating high quality sets.

4.2 Methodology

Guidelines: We developed annotation guidelines through three rounds of pilot annotations in which annotators labeled a set of messages and discussed all disagreements. We instructed annotators to link each message to the one or more messages it is a response to. If a message started a new conversation it was linked to itself. We also described a series of subtle cases, using one to three examples to tease out differences. These included when a question is repeated, when a user responds multiple times, interjections, etc. For our full guidelines, see the supplementary material. All annotations were performed using SLATE (Kummerfeld, 2019), a custom-built tool with features designed specifically for this task.[5]

Adjudication: Table 1 shows the number of annotators for each subset of our data. For the development, test, out-of-domain data, and a small set of the training data, we labeled each sample multiple times and then resolved all disagreements in an adjudication step. During adjudication, there was no indication of who had given which annotation, and there was the option to choose a different annotation entirely. In order to maximize the volume annotated, we did not perform adjudication for most of the training data. Also, the 18,924 training message set initially only had 100 messages of context per sample, and we later added another 900 lines and checked every message that was not a reply to see if it was a response to something in the additional context.

Annotators: The annotators were all fluent English speakers with a background in computer science (necessary to understand the technical content): a postdoc, a master's student, and three CS undergraduates. All adjudication was performed by the postdoc, who is a native English speaker.

Time: Annotations took between 7 and 11 seconds per message depending on the complexity of the discussion, and adjudication took 5 seconds per message. Overall, we spent approximately 240 hours on annotation and 15 hours on adjudication.

```
    [21:29] <MOUD> that reminds me... how can I use
    CTRL+C/V on terminal?
    [21:29] <MonkeyDust> MOUD ctrl ins pasts
    [21:29] <nacc> MOUD: it depends on your
    terminal application, in gnome-terminal ...
->  [21:30] <MOUD> -.-

    [17:35] <Moae> i have to remove LCDproc ...
    [17:38] <Madsy> Moae: sudo make uninstall &&
    make clean? :-)
    [17:39] <Madsy> Open the makefile and see what
    the targets are.
->  [17:40] <Madsy> Moae: Don't message people in
    private please. It's ...
    [17:42] <Moae> Madsy: sorry
    [17:42] <Moae> Madsy where i have to launch the
    command?
```

Figure 2: Examples of annotation ambiguity. Top: The message from MOUD could be a response to either nacc or MonkeyDust. Bottom: The message from Madsy could be part of this conversation or a separate exchange between the same users.

4.3 Annotation Quality

Our annotations define two levels of structure: (1) links between pairs of messages, and (2) sets of messages, where each set is one conversation. Annotators label (1), from which (2) can be inferred. Table 2 presents inter-annotator agreement measures for both cases. These are measured in the standard manner, by comparing the labels from different annotators on the same data. We also include measurements for annotations in prior work.

Figure 2 shows ambiguous examples from our data to provide some intuition for the source of disagreements. In both examples the disagreement involves one link, but the conversation structure in the second case is substantially changed. Some disagreements in our data are mistakes, where one annotation is clearly incorrect, and some are ambiguous cases, such as these. In Channel Two, we also see mistakes and ambiguous cases, including a particularly long discussion about a user's financial difficulties that could be divided in multiple ways (also noted by Elsner and Charniak (2008)).

Graphs: We measure agreement on the graph structure annotation using Cohen (1960)'s κ. This measure of inter-rater reliability corrects for chance agreement, accounting for the class imbalance between linked and not-linked pairs.

Values are in the good agreement range proposed by Altman (1990), and slightly higher than for Mehri and Carenini (2017)'s annotations. Results are not shown for Elsner and Charniak (2008) because they did not annotate graphs.

[5] https://jkk.name/slate

Conversations: We consider three metrics:[6]

(1) Variation of Information (VI, Meila, 2007). A measure of information gained or lost when going from one clustering to another. It is the sum of conditional entropies $H(Y|X) + H(X|Y)$, where X and Y are clusterings of the same set of items. We consider a scaled version, using the bound for n items that $\text{VI}(X;Y) \leq \log(n)$, and present $1-\text{VI}$ so that larger values are better.

(2) One-to-One Overlap (1-1, Elsner and Charniak, 2008). Percentage overlap when conversations from two annotations are optimally paired up using the max-flow algorithm. We follow Mehri and Carenini (2017) and keep system messages.

(3) Exact Match F_1. Calculated using the number of perfectly matching conversations, excluding conversations with only one message (mostly system messages). This is an extremely challenging metric. We include it because it is easy to understand and it directly measures a desired value (perfectly extracted conversations).

Our scores are higher in 4 cases and lower in 5. Interestingly, while κ was higher for us than Mehri and Carenini (2017), our scores for conversations are lower. This is possible because a single link can merge two conversations, meaning a single disagreement in links can cause a major difference in conversations. This may reflect the fact that our annotation guide was developed for the Ubuntu channel, which differs in conversation style from the Channel Two data. Manually comparing the annotations, there was no clear differences in the types of disagreements.

Agreement is lower on the Channel Two data, particularly on its test set. From this we conclude that there is substantial variation in the difficulty of conversation disentanglement across datasets.[7]

5 Evaluating Disentanglement Quality

In this section, we propose new simple disentanglement models that perform better than prior methods, and re-examine prior work. The models we consider are:

Previous: Each message is linked to the most recent non-system message before it.

		Graph	Conversation		
Data		κ	VI	1-1	F_1
Train (subset)		0.71	94.2	85.0	52.5
Dev		0.72	94.0	83.8	42.9
Test		0.74	95.0	83.8	49.5
Channel Two		0.72	90.4	75.9	28.2
Subparts of Channel Two					
Pilot	This work	0.68	90.9	82.4	43.5
	Elsner (2008)	-	94.2	90.0	40.7
Dev	This work	0.74	92.2	81.7	27.5
Mehri	This work	0.73	86.2	71.9	22.2
	Mehri (2017)	0.67	91.3	80.7	38.7
Test	This work	0.73	84.3	66.5	23.8
	Elsner (2008)	-	80.8	62.4	20.6

Table 2: Inter-annotator agreement for graphs (κ) and conversations (1-1, VI, F_1). Our annotations are comparable to prior work, and κ is in the good agreement range proposed by Altman (1990). We also adjudicated all disagreements to improve quality.

Lowe et al. (2017): A heuristic based on time differences and identifying directed messages.

Elsner and Charniak (2008): A linear pairwise scoring model in which each message is linked to the highest scoring previous message, or none if all scores are below zero.

Linear: Our linear ranking model that scores potential antecedents using a feature-based model based on properties such as time, directedness, word overlap, and context.

Feedforward (FF): Our feedforward model with the same features as the linear model, plus a sentence embedding calculated using an average of vectors from GloVe (Pennington et al., 2014).

Union: Run 10 FF models trained with different random seeds and combine their output by keeping all edges predicted.

Vote: Run 10 FF models and combine output by keeping the edges they all agree on. Link messages with no agreed antecedent to themselves.

Intersect: Conversations that 10 FF models agree on, and other messages as singleton conversations.

For Channel Two we also compare to Wang and Oard (2009) and Mehri and Carenini (2017), but their code was unavailable, preventing evaluation on our data. We exclude Jiang et al. (2018) as they substantially modified the dataset. For details of models, including hyperparameters tuned on the development set, see the supplementary material.

[6] Metrics such as Cohen's κ and Krippendorff's α are not applicable to conversations because there is no clear mapping from one set of conversations to another.

[7] Riou et al. (2015) also observe this, noting that their French IRC data is less entangled than Elsner's, making it possible to achieve an agreement level of 0.95.

System	P	R	F
Previous	35.7*	34.4*	35.0*
Linear	64.7	62.3	63.5
Feedforward	73.7*	71.0*	72.3*
x10 union	64.3	**79.7***	71.2*
x10 vote	**74.9***	72.2*	**73.5***

Table 3: *Graph* results on the Ubuntu test set. * indicates a significant difference at the 0.01 level compared to Linear.

System	VI	1-1	P	R	F
Previous	66.1	27.6	0.0	0.0	0.0
Linear	88.9	69.5	19.3	24.9	21.8
Feedforward	91.3	75.6	34.6	38.0	36.2
x10 union	86.2	62.5	40.4	28.5	33.4
x10 vote	**91.5**	**76.0**	36.3	**39.7**	**38.0**
x10 intersect	69.3	26.6	**67.0**	21.1	32.1
Lowe (2017)	80.6	53.7	10.8	7.6	8.9
Elsner (2008)	82.1	51.4	12.1	21.5	15.5

Table 4: *Conversation* results on the Ubuntu test set. Our new model is substantially better than prior work. Significance is not measured as we are unaware of methods for set structured data.

Training Condition	Graph-F	Conv-F
Standard	72.3 (0.4)	36.2 (1.7)
No context	72.3 (0.2)	37.6 (1.6)
1k random msg	63.0* (0.4)	21.0 (2.3)
2x 500 msg samples	61.4* (1.8)	20.4 (3.2)

Table 5: Performance with different training conditions on the Ubuntu test set. For Graph-F, * indicates a significant difference at the 0.01 level compared to Standard. Results are averages over 10 runs, varying the data and random seeds. The standard deviation is shown in parentheses.

5.1 Results

Graphs: Table 3 presents precision, recall, and F-score over links. Our models perform much better than the baseline. As we would expect, vote has higher precision, while union has higher recall. Vote has higher recall than a single feedforward model because it identifies more of the self-link cases (its default when there is no agreement).

Conversations: Table 4 presents results on the metrics defined in Section 4.3. There are three regions of performance. First, the baseline has consistently low scores since it forms a single conversation containing all messages. Second, Elsner and Charniak (2008) and Lowe et al. (2017) perform similarly, with one doing better on VI and the other on 1-1, though Elsner and Charniak (2008) do consistently better across the exact conversation extraction metrics. Third, our methods do best, with x10 vote best in all cases except precision, where the intersect approach is much better.

Dataset Variations: Table 5 shows results for the feedforward model with several modifications to the training set, designed to test corpus design decisions. Removing context does not substantially impact results. Decreasing the data size to match Elsner and Charniak (2008)'s training set leads to worse results, both if the sentences are from diverse contexts (3rd row), and if they are from just two contexts (bottom row). We also see a substantial increase in the standard deviation when only two samples are used, indicating that performance is not robust when the data is not widely sampled.

5.2 Channel Two Results

For channel Two, we consider two annotations of the same underlying text: ours and Elsner and Charniak (2008)'s. To compare with prior work, we use the metrics defined by Shen et al. (2006, Shen) and Elsner and Charniak (2008, Loc).[8] We do not use these for our data as they have been superseded by more rigorously studied metrics (VI for Shen) or make strong assumptions about the data (Loc). We do not evaluate on graphs because Elsner and Charniak (2008)'s annotations do not include them. This also prevents us from training our method on their data.

Model Comparison: For Elsner's annotations (top section of Table 6), their approach remains the most effective with just Channel Two data. However, training on our Ubuntu data, treating Channel Two as an out-of-domain sample, yields substantially higher performance on two metrics and comparable performance on the third. On our annotations (bottom section), we see the same trend. In both cases, the heuristic from Lowe et al. (2015, 2017) performs poorly. We suspect our model trained only on Channel Two data is overfitting,

[8] Loc is a Rand index that only counts messages less than 3 apart. Shen calculates the F-score for each gold-system conversation pair, finds the max for each gold conversation, and averages weighted by the size of the gold conversation (this allows a predicted conversation to match to zero, one, or multiple gold conversations). Following Wang and Oard (2009) and Mehri and Carenini (2017), we include system messages in evaluation. We also checked our metric implementations by removing system messages and calculating results for Elsner and Charniak (2008)'s output.

Test	Train	System	1-1	Loc	Shen
Elsner	Ch 2 (Elsner)	Elsner (2008)	53.1	81.9	55.1
	Ch 2 (Elsner)	Wang (2009)	47.0	75.1	52.8
	Ch 2 (Ours)	Elsner (2008)	51.1	78.0	53.9
	Ch 2 (Ours)	Feedforward	52.1	77.8	53.8
	Multiple	Mehri (2017)	55.2	78.6	56.6
	n/a	Lowe (2017)	45.1	73.8	51.8
	Ubuntu	Feedforward	**57.5**	**82.0**	**60.5**
Ours	Ch 2 (Elsner)	Elsner (2008)	54.0	81.2	56.3
	Ch 2 (Ours)	Elsner (2008)	59.7	80.8	63.0
	Ch 2 (Ours)	Feedforward	57.7	80.3	59.8
	n/a	Lowe (2017)	43.4	67.9	50.7
	Ubuntu	Feedforward	**62.8**	**84.3**	**66.6**

Table 6: Results for different annotations of Channel Two. The **best result** is bold, and the best result with only Channel Two data is underlined.

as the graph F-score on the training data is 94, whereas on the Ubuntu data it is 80.

Data Comparison: Comparing the same models in the top and bottom section, scores are consistently higher for our annotations, except for the Lowe et al. (2015, 2017) heuristic. Comparing the annotations, we find that their annotators identified between 250 and 328 conversations (mean 281), while we identify 257. Beyond this difference it is hard to identify consistent variations in the annotations. Another difference is the nature of the evaluation. On Elsner's data, evaluation is performed by measuring relative to each annotators labels and averaging the scores. On our data, we adjudicated the annotations, providing a single gold standard. Evaluating our Channel-Two-trained Feedforward model on our two pre-adjudication annotations and averaging scores, the results are lower by 3.1, 1.8, and 4.3 on 1-1, Loc and Shen respectively. This suggests that our adjudication process removes annotator mistakes that introduce noise into the evaluation.

5.3 Evaluating Lowe et al. (2015, 2017)

The previous section showed that only 10.8% of the conversations extracted by the heuristic in Lowe et al. (2015, 2017) are correct (P in Table 4). We focus on precision because the primary use of their method has been to extract conversations to train and test dialogue systems, which will be impacted by errors in the conversations. Recall errors (measuring missed conversations) are not as serious a problem because the Ubuntu chat logs are so large that even with low recall a large number of conversations will still be extracted.

```
Missed [02:06] <TheBuntu> in virtualbox...  win7 in
       VM...  i have an ntfs partition..  How do i
       access that partition in VM ?
       [02:06] <L1nuxRules> share it with the vm
       [02:08] <L1nuxRules> anywy this is ubuntu so
       windows &> /duv/null
       [02:09] <L1nuxRules> dev*
Extra  [02:11] <L1nuxRules> it shouldnt unless
       theres depency issues
       [02:11] <TheBuntu> L1nuxRules:  how do i
       share with the vm...  i dont see VM in share
Missed [02:12] <L1nuxRules> buntu if its virtuasl
       box click on setttings > shared folders
Missed [02:13] <TheBuntu> ok
```

Figure 3: An example conversation extracted by the heuristic from Lowe et al. (2015, 2017) with the messages it misses and the one it incorrectly includes.

Additional Metrics: First, we must check this is not an artifact of our test set. On our development set, P, R, and F are slightly higher (11.6, 8.1 and 9.5), but VI and 1-1 are slightly lower (80.0 and 51.7). We can also measure performance as the distribution of scores over all of the samples we annotated. The average precision was 10, and varied from 0 to 50, with 19% of cases at 0 and 95% below 23. To avoid the possibility that we made a mistake running their code, we also considered evaluating their released conversations. On the data that overlapped with our annotations, the precision was 9%. These results indicate that the test set performance is not an aberration: the heuristic's results are consistently low, with only about 10% of output conversations completely right.

Error Types: Figure 3 shows an example heuristic output with several types of errors. The initial question was missed, as was the final resolution, and in the middle there is a message from a separate conversation. 67% of conversations were a subset of a true conversation (ie., only missed messages), and 3% were a superset of a true conversation (ie., only had extra messages). The subset cases were missing 1-187 messages (missing 56% of the conversation on average) and the superset cases had 1-3 extra messages (an extra 31% of the conversation on average). The first message is particularly important because it is usually the question being resolved. In 47% of cases the first message is not the true start of a conversation.

It is important to note that the dialogue task the conversations were intended for only uses a prefix of each conversation. For this purpose, missing the end of a conversation is not a problem. In 9% of cases, the conversation is a true prefix of a gold conversation. Combined with the exact match cases, that means 20% of the conversations are accurate as used in the next utterance selection task. A further 9% of cases are a continuous

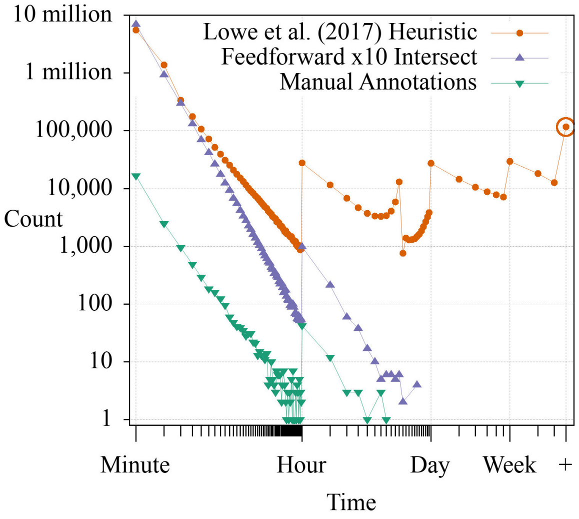


Figure 4: Time between consecutive messages in conversations. Jumps are at points when the scale shifts as indicated on the x-axis. The circled upper right point is the sum over all larger values, indicating that messages weeks apart are often in the same conversation.

chunk of a conversation, but missing one or more messages at the start.

Long Distance Links: One issue we observed is that conversations often spanned days. We manually inspected a random sample: 20 conversations 12 to 24 hours long, and 20 longer than 24 hours. All of the longer conversations and 17 of the shorter ones were clearly incorrect.[9] This issue is not measured in the analysis above because our samples do not span days (they are 5.5 hours long on average when including context). The original work notes this issue, but claims that it is rare. We measured the time between consecutive messages in conversations and plot the frequency of each value in Figure 4.[10] The figure indicates that the conversations often extend over days, or even more than a month apart (note the point in the top-right corner). In contrast, our annotations rarely contain links beyond an hour, and the output of our model rarely contains links longer than 2 hours.

Causes: To investigate possible reasons for these issues, we measured several properties of our data to test assumptions in the heuristic. First, the heuristic assumes if all directed messages from a user are in one conversation, all undirected messages from the user are in the same conversation. We find this is true 52.2% of the time. Second, it assumes that it is rare for two people to respond to an initial question. In our data, of the messages that start a conversation and receive a response, 37.7% receive multiple responses. Third, that a directed message can start a conversation, which we find in 6.8% of cases. Fourth, that the first response to a question is within 3 minutes, which we find is true in 94.8% of conversations. Overall, these assumptions have mixed support from our data, which may be why the heuristic produces so few accurate conversations.

Model	Test	Train	MRR	R@1	R@5
DE	Lowe	Lowe	0.75	0.61	0.94
		Ours	0.63	0.45	0.90
	Ours	Lowe	0.72	0.57	0.93
		Ours	0.76	0.63	0.94
ESIM	Lowe	Lowe	0.82	0.72	0.97
		Ours	0.69	0.53	0.92
	Ours	Lowe	0.78	0.67	0.95
		Ours	0.83	0.74	0.97

Table 7: Next utterance prediction results, with various models and training data variations. The decrease in performance when training on one set and testing on the other suggests they differ in content.

Dialogue Modeling: Most of the work building on Lowe et al. (2017) uses the conversations to train and evaluate dialogue systems. To see the impact on downstream work, we constructed a next utterance selection task as described in their work, disentangling the entire `#Ubuntu` logs with our feedforward model. We tried two dialogue models: a dual-encoder (Lowe et al., 2017), and Enhanced Long Short-Term Memory (Chen et al., 2017b). For full details of the task and model hyperparameters, see the supplementary material.

Table 7 show results when varying the training and test datasets. Training and testing on the same dataset leads to higher performance than training on one and testing on the other. This is true even though the heuristic data contains nine times as many training conversations. This is evidence that our conversations are fundamentally different despite being derived from the same resource and filtered in the same way. This indicates that our changes lead to quantitatively different downstream models. Fortunately, the relative performance of the two models remains consistent across the two datasets.

[9] The exceptions were two cases where a user thanked another user for their help the previous day, and one case where a user asked if another user ended up resolving their question.

[10] In 68,002 conversations there was a negative time difference because a message was out of order. To resolve this, we sorted the messages in each conversation by timestamp.

5.4 Re-Examining Disentanglement Research

Using our data we also investigate other assumptions made in prior work. The scale of our data provides a more robust test of these ideas.

Number of samples: Table 1 shows that all prior work with available data has considered a small number of samples. In Table 5, we saw that training on less diverse data samples led to models that performed worse and with higher variance. We can also investigate this by looking at performance on the different samples in our test set. The difficulty of samples varies considerably, with the F-score of our model varying from 11 to 40 and annotator agreement scores before adjudication varying from 0.65 to 0.78. The model performance and agreement levels are also strongly correlated, with a Spearman's rank correlation of 0.77. This demonstrates the importance of evaluating on data from more than one point in time to get a robust estimate of performance.

How far apart consecutive messages in a conversation are: Elsner and Charniak (2008) and Mehri and Carenini (2017) use a limit of 129 seconds, Jiang et al. (2018) limit to within 1 hour, Guo et al. (2017) limit to within 8 messages, and we limit to within 100 messages. Figure 4 shows the distribution of time differences in our conversations. 94.9% are within 2 minutes, and almost all are within an hour. 88.3% are 8 messages or less apart, and 99.4% are 100 or less apart. This suggests that the lower limits in prior work are too low. However, in Channel Two, 98% of messages are within 2 minutes, suggesting this property is channel and sample dependent.

Concurrent conversations: Adams and Martell (2008) forced annotators to label at most 3 conversations, while Jiang et al. (2018) remove conversations to ensure there are no more than 10 at once. We find there are 3 or fewer 46.4% of the time and 10 or fewer 97.3% of the time (where time is in terms of messages, not minutes, and we ignore system messages), Presumably the annotators in Adams and Martell (2008) would have proposed changes if the 3 conversation limit was problematic, suggesting that their data is less entangled than ours.

Conversation and message length: Adams and Martell (2008) annotate blocks of 200 messages. If such a limit applied to our data, 13.7% of conversations would not finish before the cutoff point. This suggests that their conversations are typically shorter, which is consistent with the previous conclusion that their conversations are less entangled. Jiang et al. (2018) remove conversations with fewer than 10 messages, describing them as outliers, and remove messages shorter than 5 words, arguing that they were not part of real conversations. Not counting conversations with only system messages, 83.4% of our conversations have fewer than 10 messages, 40.8% of which have multiple authors. 88.5% of messages with less than 5 words are in conversations with more than one author. These values suggest that these messages and conversations are real and not outliers.

Overall: This analysis indicates that working from a small number of samples can lead to major bias in system design for disentanglement. There is substantial variation across channels, and across time within a single channel.

6 Conclusion

Conversation disentanglement has been understudied because of a lack of public, annotated datasets. We introduce a new corpus that is larger and more diverse than any prior corpus, and the first to include context and adjudicated annotations. Using our data, we perform the first empirical analysis of Lowe et al. (2015, 2017)'s widely used data, finding that only 20% of the conversations their method produces are true prefixes of conversations. The models we develop have already enabled new directions in dialogue research, providing disentangled conversations for DSTC 7 track 1 (Gunasekara et al., 2019; Yoshino et al., 2018) and will be used in DSTC 8. We also show that diversity is particularly important for the development of robust models. This work fills a key gap that has limited research, providing a new opportunity for understanding synchronous multiparty conversation online.

Acknowledgements

We would like to thank Jacob Andreas, Greg Durrett, Will Radford, Ryan Lowe, and Glen Pink for helpful feedback on earlier drafts of this paper and the anonymous reviewers for their helpful suggestions. This material is based in part on work supported by IBM as part of the Sapphire Project at the University of Michigan. Any opinions, findings, conclusions or recommendations expressed above do not necessarily reflect the views of IBM.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Page H. Adams and Craig H. Martell. 2008. Topic Detection and Extraction in Chat. In *2008 IEEE International Conference on Semantic Computing*.

Douglas G Altman. 1990. *Practical statistics for medical research*. CRC press.

Erik Aumayr, Jeffrey Chan, and Conor Hayes. 2011. Reconstruction of threaded conversations in online discussion forums. In *International AAAI Conference on Web and Social Media*.

Ali Balali, Hesham Faili, and Masoud Asadpour. 2014. A Supervised Approach to Predict the Hierarchical Structure of Conversation Threads for Comments. *The Scientific World Journal*.

Ali Balali, Hesham Faili, Masoud Asadpour, and Mostafa Dehghani. 2013. A Supervised Approach for Reconstructing Thread Structure in Comments on Blogs and Online News Agencies. *Computacion y Sistemas*, 17(2):207–217.

Jun Chen, Chaokun Wang, Heran Lin, Weiping Wang, Zhipeng Cai, and Jianmin Wang. 2017a. *Learning the Structures of Online Asynchronous Conversations*, volume 10177 of *Lecture Notes in Computer Science*. Springer.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Giacomo Domeniconi, Konstantinos Semertzidis, Vanessa Lopez, Elizabeth M. Daly, Spyros Kotoulas, and Gianluca Moro. 2016. A Novel Method for Unsupervised and Supervised Conversational Message Thread Detection. In *Proceedings of the 5th International Conference on Data Management Technologies and Applications - Volume 1: DATA,*.

Wenchao Du, Pascal Poupart, and Wei Xu. 2017. Discovering Conversational Dependencies between Messages in Dialogs. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Andrei Dulceanu. 2016. Recovering implicit thread structure in chat conversations. *Revista Romana de Interactiune Om-Calculator*, 9:217–232.

Micha Elsner and Eugene Charniak. 2008. You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement. In *Proceedings of ACL-08: HLT*.

Micha Elsner and Eugene Charniak. 2010. Disentangling Chat. *Computational Linguistics*, 36(3):389–409.

Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Micha Elsner and Warren Schudy. 2009. Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*.

Chulaka Gunasekara, Jonathan K. Kummerfeld, Lazaros Polymenakos, , and Walter S. Lasecki. 2019. Dstc7 task 1: Noetic end-to-end response selection. In *7th Edition of the Dialog System Technology Challenges at AAAI 2019*.

Gaoyang Guo, Chaokun Wang, Jun Chen, and Pengcheng Ge. 2017. Who Is Answering to Whom? Finding "Reply-To" Relations in Group Chats with Long Short-Term Memory Networks. In *International Conference on Emerging Databases (EDB'17)*.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and Linking Web Forum Posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*.

Jonathan K. Kummerfeld. 2019. Slate: A super-lightweight annotation tool for experts. In *Proceedings of ACL 2019, System Demonstrations*.

Walter S. Lasecki, Ece Kamar, and Dan Bohus. 2013. Conversations in the crowd: Collecting data for task-oriented dialog learning. In *Proceedings of the Human Computation Workshop on Scaling Speech, Language Understanding and Dialogue through Crowdsourcing*.

Annie Louis and Shay B. Cohen. 2015. Conversation Trees: A Grammar Model for Topic Structure in Forums. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training End-to-End Dialogue Systems with the Ubuntu Dialogue Corpus. *Dialogue & Discourse*, 8(1).

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2012. Hierarchical Conversation Structure Prediction in Multi-Party Chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Marina Meila. 2007. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthieu Riou, Soufian Salim, and Nicolas Hernandez. 2015. Using discursive information to disentangle French language chat. In *NLP4CMC 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media at GSCL Conference*.

Anna Schuth, Maarten Marx, and Maarten de Rijke. 2007. Extracting the discussion structure in comments on news-articles. In *Proceedings of the 9th annual ACM international workshop on Web information and data management*.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread Detection in Dynamic Text Message Streams. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

David Uthus and David Aha. 2013a. Detecting Bot-Answerable Questions in Ubuntu Chat. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

David Uthus and David Aha. 2013b. Extending word highlighting in multiparticipant chat. In *Florida Artificial Intelligence Research Society Conference*.

David C. Uthus and David W. Aha. 2013c. The Ubuntu Chat Corpus for Multiparticipant Chat Analysis. In *Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium*.

Hongning Wang, Chi Wang, ChengXiang Zhai, and Jiawei Han. 2011a. Learning Online Discussion Structures by Conditional Random Fields. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011b. Predicting thread discourse structure over technical web forums. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 13–25.

Lidan Wang and Douglas W. Oard. 2009. Context-based Message Expansion for Disentanglement of Interleaved Text Conversations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Yi-Chia Wang, Mahesh Joshi, William Cohen, and Carolyn Rosé. 2008. Recovering Implicit Thread Structure in Newsgroup Style Conversations. In *Proceedings of the International Conference on Weblogs and Social Media*.

Yi-Chia Wang and Carolyn P. Rosé. 2010. Making Conversational Structure Explicit: Identification of Initiation-response Pairs within Online Discussions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S. Lasecki, Jonathan K. Kummerfeld, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Xiang Gao, Huda Alamari, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2018. Dialog system technology challenge 7. In *NeurIPS Workshop: The 2nd Conversational AI: "Today's Practice and Tomorrow's Potential"*.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing Online Discussion Using Coarse Discourse Sequences. In *11th AAAI International Conference on Web and Social Media (ICWSM)*.